Letter

Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*

Jennifer S. Hawkins,[1] HyeRan Kim,[2] John D. Nason,[1] Rod A. Wing,[2] and Jonathan F. Wendel[1,3]

[1]*Iowa State University, Department of Ecology, Evolution and Organismal Biology, Ames, Iowa 50011, USA;* [2]*University of Arizona, Department of Plant Sciences, Arizona Genomics Institute, Tucson, Arizona 85721, USA*

The DNA content of eukaryotic nuclei (C-value) varies ~200,000-fold, but there is only a ~20-fold variation in the number of protein-coding genes. Hence, most C-value variation is ascribed to the repetitive fraction, although little is known about the evolutionary dynamics of the specific components that lead to genome size variation. To understand the modes and mechanisms that underlie variation in genome composition, we generated sequence data from whole genome shotgun (WGS) libraries for three representative diploid ($n = 13$) members of *Gossypium* that vary in genome size from 880 to 2460 Mb (1C) and from a phylogenetic outgroup, *Gossypioides kirkii*, with an estimated genome size of 588 Mb. Copy number estimates including all dispersed repetitive sequences indicate that 40%–65% of each genome is composed of transposable elements. Inspection of individual sequence types revealed differential, lineage-specific expansion of various families of transposable elements among the different plant lineages. *Copia*-like retrotransposable element sequences have differentially accumulated in the *Gossypium* species with the smallest genome, *G. raimondii*, while *gypsy*-like sequences have proliferated in the lineages with larger genomes. Phylogenetic analyses demonstrated a pattern of lineage-specific amplification of particular subfamilies of retrotransposons within each species studied. One particular group of *gypsy*-like retrotransposon sequences, *Gorge3* (*Gossypium* retrotransposable *gypsy*-like element), appears to have undergone a massive proliferation in two plant lineages, accounting for a major fraction of genome-size change. Like maize, *Gossypium* has undergone a threefold increase in genome size due to the accumulation of LTR retrotransposons over the 5–10 Myr since its origin.

[The sequence data described in this paper have been submitted to the GSS Division of GenBank under accessions DX390732–DX406528.]

Genomes of eukaryotic organisms vary over 200,000-fold in size, ranging from 2.8 Mb in *Encephalitozoon cuniculi* (Biderre et al. 1998) to >690,000 Mb in the diatom *Navicola pelliculosa* (Cavalier-Smith 1985; Li and Graur 1991). Among angiosperms, genome sizes range from ~108 Mb for *Fragaria viridis* (Bennett and Leitch 2005) to >120,000 Mb in some members of the Liliaceae (Flavell et al. 1974; Bennett and Smith 1991; Bennett and Leitch 1995, 1997; Leitch et al. 1998). Not only is wide variation in genome size common among distantly related organisms, but it also is unexceptional even among closely related species. For example, genome sizes range approximately sixfold among members of the genus *Vicia* (Chooi 1971), and ninefold within the genus *Crepsis* (Jones and Brown 1976). Some portion of this genome size variation may be ascribed to differences in gene number amplification due to gene, chromosome segment, and whole-genome duplication, as well as to gene loss (Tikhonov et al. 1999; Blanc et al. 2000; Grant et al. 2000; Ku et al. 2000; Vision et al. 2000; Wendel 2000; Bancroft 2001; Bennetzen and Ramakrishna 2002). Nevertheless, >90% of plant genes possess close homologs within other plant species, indicative of highly conserved gene content (Bennetzen 2000a).

[3]Corresponding author.
E-mail jfw@iastate.edu; fax (515) 294-1337.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5282906.

There appears to be no correlation between the amount of DNA per cell and organismal advancement or genetic complexity (Sparrow et al. 1972; Price 1988). This well-documented lack of correspondence between genome size and morphological or physiological complexity of an organism has been historically termed the "C-value paradox" (Thomas 1971). Since the discovery of non-coding DNA and its impact on genome size variation, "paradox" has been replaced by "enigma" in an attempt to more appropriately identify the topic as a "perplexing subject" made up of several independent components (Gregory 2002, 2004). It is now generally agreed that the C-value enigma can be largely explained by the differential amplification and proliferation among organisms of the repetitive fraction of the genome (Bennetzen 2000b, 2002; Kidwell 2002).

In plants, amplification and insertion of newly activated long terminal repeat (LTR) retrotransposable elements appear to be major contributors to genome size expansion. For example, ~70% of the maize nuclear genome is composed of LTR-retrotransposons (SanMiguel and Bennetzen 1998). In the span of just a few million years, the maize genome doubled in size due to transposable element (TE) activity (SanMiguel and Bennetzen 1998). These TEs are often found in nested arrangements located between "gene islands," and often are associated with centromeres (SanMiguel et al. 1996). To date, little is known regarding the extent to which various TEs contribute to genome size variation or how TE types are distributed among closely related spe-

16:1252–1261 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

cies. Other mechanisms posited to be responsible for genome size expansion include variation in intron size (Deutsch and Long 1999), expansion of tandemly repetitive DNA sequences (Ellegren 2002; Morgante et al. 2002), segmental duplication (Blanc et al. 2000; Ku et al. 2000; Vision et al. 2000; Wendel 2000; Bancroft 2001), accumulation of pseudogenes (Zhang 2003), and transfer of organellar DNA to the nucleus (Adams and Palmer 2003; Shahmuradov et al. 2003). However, these mechanisms generally do not appear to have a large impact on genome size differences among closely related species.

Although it has been suggested that organisms may have a "one-way ticket to genomic obesity" (Bennetzen and Kellogg 1997), it is possible that differences in genome size are not only the outcome of an organism's tolerance for accrual of nongenic DNA, but also its efficiency in removal of non-essential DNA (Petrov and Hartl 1997; Petrov et al. 2000; Petrov 2002a; Wendel et al. 2002b). Many organisms with smaller genomes are striking in their relatively small proportion of nongenic DNA. Evidence of a deletional bias among organisms with smaller versus larger genomes (Bennett and Leitch 1997; Petrov and Hartl 1997; Kirik et al. 2000; Petrov et al. 2000) has led to the "mutational equilibrium model" of DNA loss (Petrov 2002b). Other suggested mechanisms of DNA loss include unequal intrastrand homologous recombination between two tandem repeats in the same orientation, such as the LTRs of retrotransposable elements (Shepherd et al. 1984; SanMiguel et al. 1996; Chen et al. 1998; Vicient et al. 1999; Bennetzen 2002), illegitimate recombination (Devos et al. 2002; Wicker et al. 2003; Ma et al. 2004; Bennetzen et al. 2005), and double-stranded break repair (Kirik et al. 2000; Orel and Puchta 2003; Filkowski et al. 2004).

To effectively study genome size evolution from a phylogenetic perspective, it is necessary to exploit a system in which the closely related species vary widely in genome size and for whom phylogenetic relationships are well understood. A good example in this respect is the monophyletic genus *Gossypium* (Malvaceae), which is composed of ~50 species of small trees and shrubs with an aggregate distribution that encompasses many tropical and subtropical semi-arid regions of the world (Fryxell 1992; Seelanan et al. 1997; Cronn et al. 2002; Wendel and Cronn 2003). Diploid members of the genus are divided into eight groups based on chromosome pairing behavior and fertility in interspecific hybrids (Beasley 1941; Endrizzi et al. 1985). All diploid members of the genus have 13 chromosomes, yet genome sizes range approximately threefold, from a median estimate of 885 Mb per haploid nucleus in the American D-genome species, to 2572 Mb per haploid nucleus in the Australian K-genome species (Fig. 1; Hendrix and Stewart 2005). An even larger range in genome size is observed in the tribe to which *Gossypium* belongs (the *Gossypieae*), from only 590 Mb in *Gossypioides kirkii* and *Kokia drynarioides* to 4018 Mb in *Thespesia populnea* (Wendel et al. 2002b). The wide range in genome size observed across closely related diploid species and the well-established phylogeny makes *Gossypium* an excellent system for the study of genome size evolution.

To better appreciate the relevance of genome size variation to organismal fitness and evolution, it is first necessary to enhance our understanding of the quantity and quality of the genomic components that distinguish two or more genomes, as well as the modes and mechanisms by which these differences arise. This insight may derive from comparative sequence analysis of specific genomic regions or from using more global approaches. An example of the former is the recent study by Grover et al. (2004), who compared ~104 kb of aligned sequence surrounding the *CesA1* gene from the D- and A- genomes of tetraploid cotton. In this case, both gene content and intergenic regions were largely conserved, and hence there was no evidence of the mechanisms responsible for the twofold size variation that characterizes these genomes. Here, we employ the second approach, utilizing whole genome shotgun (WGS) libraries constructed for three members of *Gossypium* that range threefold in genome size, and one outgroup species, *Gossypioidies kirkii*. Copy number estimates for several *Gossypium* transposable elements suggest that different types of repetitive sequences have accumulated at different rates in different plant lineages. Additionally, the results suggest that different families within a repetitive sequence type proliferate differentially. Indeed, the major fraction of the genome size variation observed in *Gossypium* is largely due to recent, lineage-specific amplification of one particular group of *gypsy*-like retrotransposon sequences, *Gorge3* (*Gossypium* retrotransposable gypsy-like element), within the larger-genome *Gossypium* species.

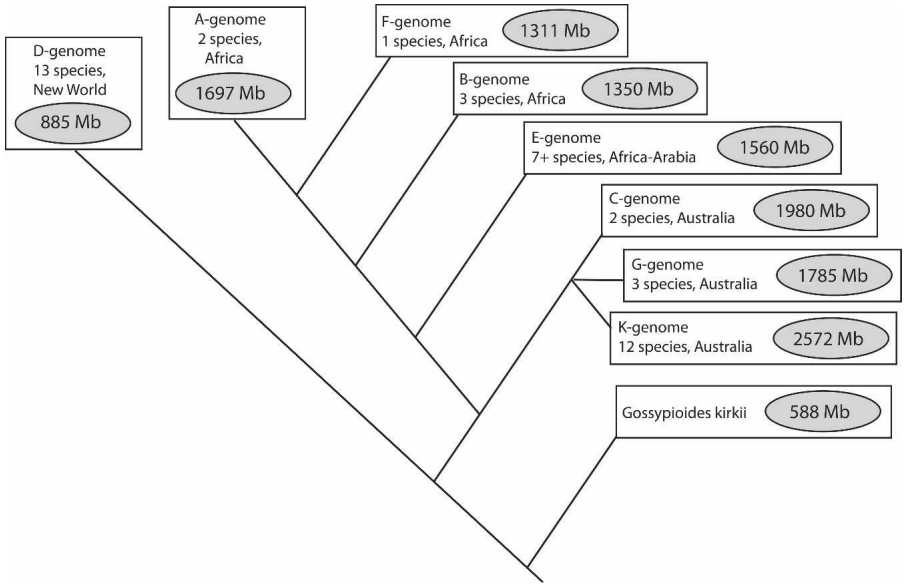

Figure 1. Evolutionary relationships among diploid members of *Gossypium*. *Gossypium* is a monophyletic genus composed of ~50 species that are widely distributed throughout many tropical and subtropical regions. Diploid species have a haploid complement of 13 chromosomes. *Gossypium* is divided into eight genome groups based on cytogenetic data and level of fertility in interspecific hybrids (Endrizzi et al. 1985). Multiple molecular data sets support the phylogenetic relationships indicated, including the outgroup relationship of *Gossypioides kirkii* (Wendel and Albert 1992; Seelanan et al. 1997; Small et al. 1998, 1999). Despite conservation of chromosome number among the diploids, genome size varies threefold, from an average of 885 Mb in the New World D-genome species to an average of 2576 Mb in the Australian K-genome species (Hendrix and Stewart 2005).

Table 1. Library construction and sequencing effort for three species representing different *Gossypium* genomes and one phylogenetic outgroup

Taxon/genome group	Genome size[a] (in Mb)	No. clones in library	Successfully sequenced	Average read (bp)	% genome sequenced	No. Mb sequenced
Gossypioides kirkii Outgroup	588	1920	1464	753	0.19%	1.10
Gossypium raimondii D genome	880	3072	2722	770	0.24%	2.10
G. herbaceum A genome	1667	6048	4864	704	0.21%	3.42
G. exiguum K genome	2460	10368	6747	704	0.19%	4.75
					TOTAL	11.4

[a]Genome size from Wendel et al. (2002b) for *G. kirkii* and from Hendrix and Stewart (2005) for *G. raimondii*, *G. herbaceum*, and *G. exiguum*.

Results

Library construction and sequence analysis

The *Gossypioides kirkii* (outgroup), *Gossypium raimondii* (D), *G. herbaceum* (A), and *G. exiguum* (K) libraries contained 1920, 3072, 6048, and 10,368 clones, of which 1464, 2722, 4864, and 6747 were successfully sequenced, respectively (Table 1). The percent of each genome sequenced (0.19%–0.24%) was determined by multiplying the number of successfully sequenced clones by the average high-quality sequencing read length, divided by the estimated genome size. Sequences were queried against GenBank using BLASTX and against each other using BLASTN, and sequences were classified as described (see Methods). All types of dispersed repetitive sequences identified via this procedure were categorized into 1) *gypsy*-like, 2) *copia*-like, 3) LINE-like, 4) *Mutator*-like, 5) *hAT*-like, 6) *En/Spm*-like, and 7) unknown repetitive sequences. *Gypsy*- and *copia*-like LTR retrotransposons, in addition to LINE-like retroposons, were abundant in all four species (Table 2; Fig. 2). Class II DNA sequences and tandem repeats were less abundant. Some classes of dispersed repetitive sequences, such as MITEs and SINEs, were not identified in the libraries. However, because of the lack of conserved domains for these two types of sequences, they may be present in *Gossypium* and unidentifiable via BLAST. Copy number estimates suggest a minimum of 44%, 54%, 52%, and 60% of the *G. kirkii*, *G. raimondii*, *G. herbaceum*, and *G. exiguum* genomes, respectively, are occupied by repetitive sequences alone.

Several conserved coding domains for diverse repetitive sequences were recovered from the WGS libraries when queried against *Arabidopsis* and *Brassica* databases. A total of 427 *gypsy*-like reverse transcriptase sequences were identified. Phylogenetic analysis of 373 of these sequences confirmed the existence of three distinct classes of *gypsy*-like retrotransposons among the four libraries identified in the initial BLAST search (Fig. 3, and see

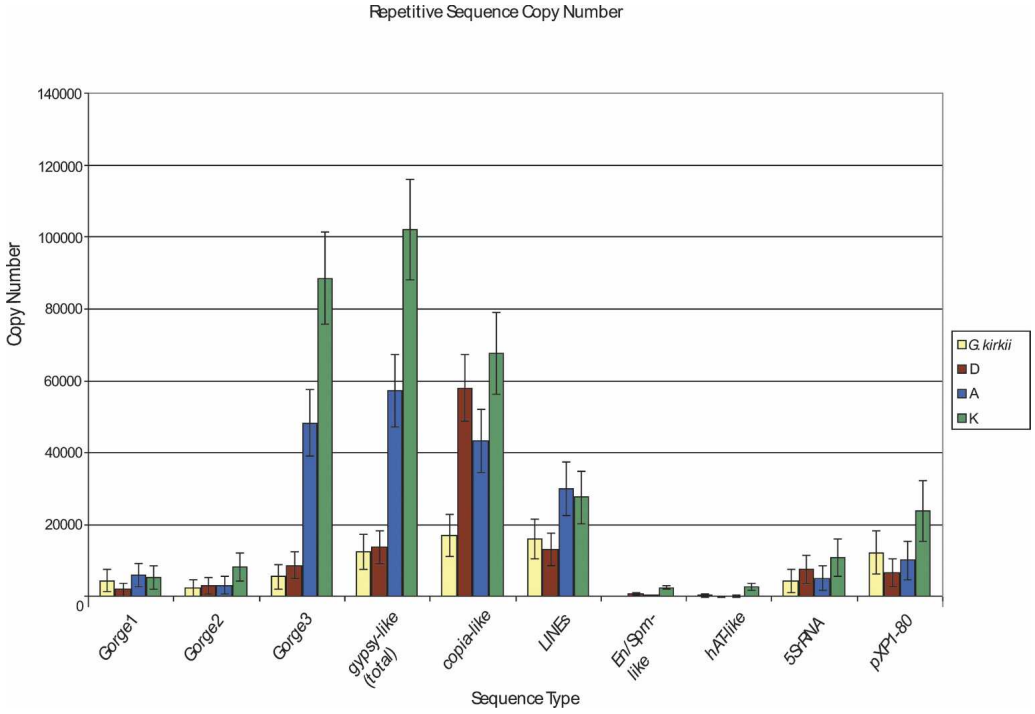

Figure 2. Copy number estimates for repetitive sequences in *Gossypium*. Copy numbers for repetitive sequences recovered in the WGS libraries were estimated as described (see Methods). The majority of repetitive sequences are LTR retrotransposons, particularly in the larger-genome species. In both the A and K genomes, massive amplification of *Gorge3 gypsy*-like sequences has occurred, contributing predominantly to genome size expansion in these two lineages. In the smallest *Gossypium* genome, *G. raimondii* (D genome), and *copia*-like sequences have proliferated and are primarily responsible for genome size expansion in this lineage. Class II sequences were less abundant and appear to contribute little to genome size evolution in the genus. Tandem repeats are approximately evenly distributed among all four species, with pXP1–80 sequences slightly elevated in *G. exiguum* (K genome).

Table 2. Repetitive element copy number and density estimates

	G. kirkii Outgroup 588 Mb	*G. raimondii* D genome 880 Mb	*G. herbaceum* A genome 1667 Mb	*G. exiguum* K genome 2460 Mb
Tandem repeats				
5SrRNA	4279 ± 3227	7675 ± 3826	5073 ± 3379	10,794 ± 5082
pXP1–80	12,264 ± 6098	6573 ± 3956	10,101 ± 5392	23,795 ± 8528
Class II transposons				
En/Spm-like	120 ± 138	835 ± 326	343 ± 216	2514 ± 602
	~0.2%	~0.9%	~0.2%	~1.0%
hAT-like	305 ± 352	81 ± 163	263 ± 304	2615 ± 986
	~0.2%	<0.1%	<0.1%	~0.4%
Class II Total	3.5 Mb	12 Mb	5 Mb	42 Mb
	<0.1%	1.0%	<0.1%	~1.4
Class I retrotransposons				
copia-like	17,006 ± 5765	57,956 ± 9300	43,181 ± 8774	67,700 ± 11,324
	9.7%–19.7%	28%–38.7%	10.7%–16.1%	11.7%–16.5%
LINE	16,006 ± 5597	13,011 ± 4502	30,000 ± 7335	27,563 ± 7271
	5.1%–10.6%	2.8%–5.7%	4.0%–6.5%	2.4%–4.1%
GORGE1 gypsy-like	4502 ± 2992	1971 ± 1762	5909 ± 3273	5319 ± 3205
	2.4%–11.9%	0.2%–3.9%	1.5%–5.2%	0.8%–3.2%
GORGE2 gypsy-like	2500 ± 2233	3154 ± 2227	3181 ± 2403	8221 ± 3983
	0.4%–7.5%	1.0%–5.7%	0.4%–3.2%	1.6%–4.7%
GORGE3 gypsy-like	5502 ± 3305	8674 ± 3683	48,181 ± 9257	88,492 ± 12,904
	3.5%–13.9%	5.3%–13.0%	22.0%–32.6%	28.8%–38.6%
Class I Total	255 Mb	465 Mb	865 Mb	1400 Mb
	42%	53%	52%	58%

below). Reverse transcriptase sequences from *copia*-like retrotransposons ($n = 71$) and LINE-like retroposons ($n = 20$) in addition to transposase sequences from *hAT*-like ($n = 2$), *Mutator*-like ($n = 1$), and *En/Spm*-like ($n = 15$) transposable elements were also retained for further analysis.

Tandem repeats were identified using *Tandem Repeat Finder* (Benson 1999). Sequences identified by *TRF* as tandemly repetitive were queried against GenBank using BLASTN in an attempt to assign sequence identity. *Gossypium* 5SrDNA sequences and a previously published *Gossypium* sequence, pXP1–80 (Zhao et al. 1998), were recovered from all four of the WGS libraries (see below).

Copy number estimates and lineage specific amplification

Tandem repeats

Sequences with high identity to previously described *Gossypium* 5SrDNA repeats were identified in all four libraries. Estimates for D- (7675 ± 3826) and A- (5073 ± 3379) genome 5SrDNA copy numbers are in agreement with previously published estimates (Cronn et al. 1996) of 4730 ± 893 for *G. raimondii* and 3415 ± 807 for *G. herbaceum*, those of the latter study being based on Southern hybridization data. Estimated copy numbers for 5SrDNA sequences among the four libraries fall well within the same 95% CI (Table 2). Several other tandem repeats were recovered. One of these tandem repeats was identified as a previously published *Gossypium* repeat, pXP1–80 (Zhao et al. 1998). This 170- to 172-bp repeat was present in all four of the WGS libraries. Similar to the 5SrDNA repeats, copy number estimates for pXP1–80 were comparable between three of the four species (*G. kirkii*—12,263 ± 6098; *G. raimondii*—6573 ± 3956; *G. herbaceum*—10,101 ± 5391) but elevated in *G. exiguum* (23,795 ± 8528). It may be that pXP1–80 is a centromere repeat, given that it is present in all of the WGS libraries, and its length is similar to those of published centromere repeats from *Arabidopsis* (178 bp), wheat (192 bp), rice (155 bp), and maize (156 bp) (Ananiev et al. 1998; Hall et al. 2003; Ito et al. 2004; Nagaki et al. 2004). Several other tandem repeats of unknown identity were identified by *TRF*. However, none of the remaining tandem repeats were shared among the WGS libraries, and all were present in low copy number.

Class II transposons

The three major superfamilies of Class II DNA transposons present in the WGS libraries are members of the *En/Spm*, *Mutator*, and *hAT* DNA transposon families. Class II sequences identified were few in number, with copy number estimates suggesting that, taken as a whole, these sequences occupy <2% of the *Gossypium* genome (Table 2). *En/Spm*-like sequences occupy <1% of the genome in each of the four species, comprising ~0.2% of the *G. kirkii* (~120 copies) and *G. herbaceum* (~343 copies) genomes, but increasing in copy number in the smallest (*G. raimondii*—0.9%, ~835 copies) and largest (*G. exiguum*—1.0%, ~2515 copies) genomes. Similarly, *hAT*-like sequences occupy <1% of the genome in each of the four species. *hAT*-like sequence comprise only 0.2% of the *G. kirkii* (~300 copies) genome, and an even smaller portion of the *G. raimondii* (0.03%, ~80 copies) and *G. herbaceum* (0.06%, ~260 copies) genomes. However, a large increase in copy number occurred in the K genome lineage (0.4%, ~2600 copies). *Mutator*-like sequences were identified in the WGS library, but because of the large range in published lengths for these sequences and the absence of a described *Mutator*-like transposon for *Gossypium*, it was not possible to estimate their copy numbers with confidence. Additionally, because of the degenerate nature of the identified Class I sequences, it is likely that there are other undetected sequences of this type in *Gossypium*. There was no evidence of MITEs, TRIMs, LARDs, or Helitrons in the WGS libraries.

Class I retrotransposons

The most highly represented group of repetitive sequences within all four WGS libraries is the Class I elements (Fig. 2). Estimated total Class I copy numbers range 4.4-fold, from 45,515 ± 9241 in *Gossypioides kirkii* to ~197,294 ± 18,935 in the K genome species, *Gossypium exiguum*. When multiplied by an

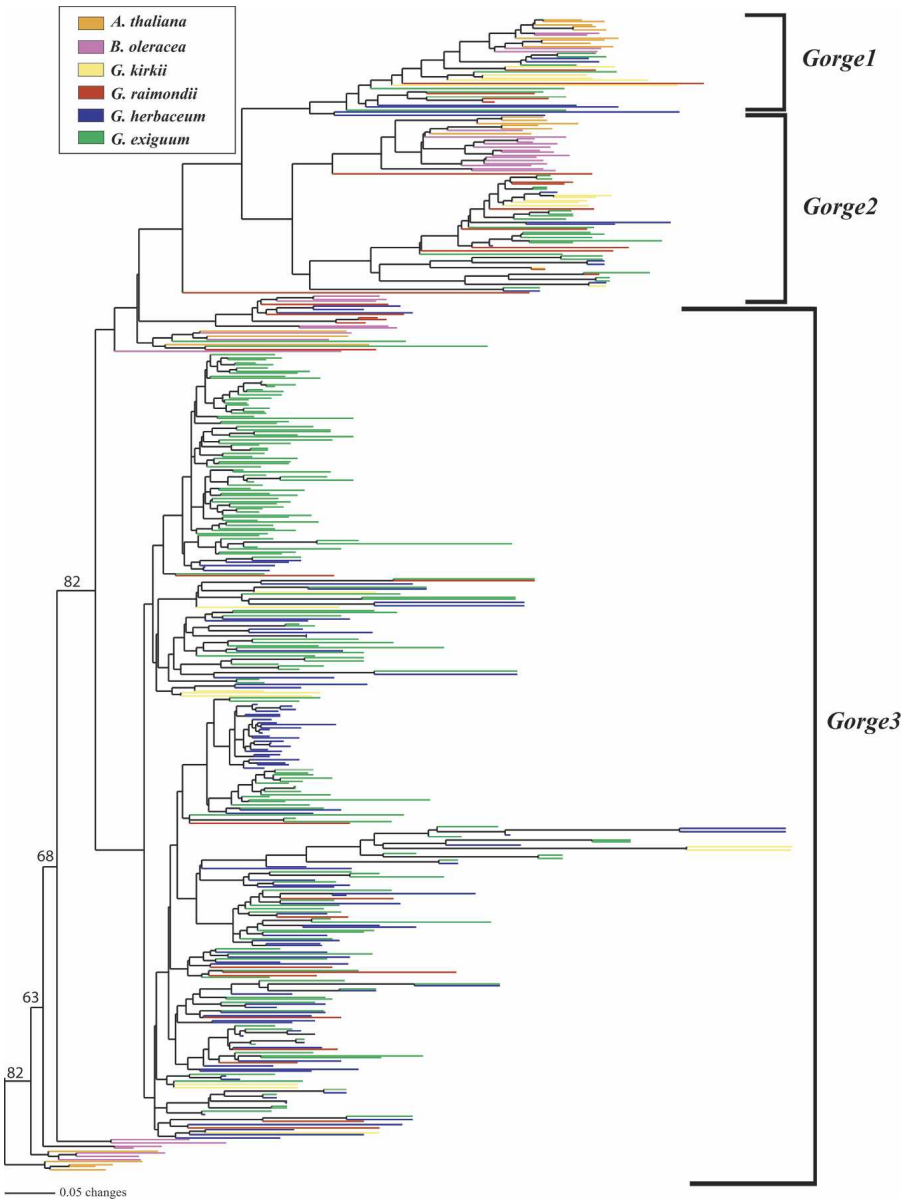

Figure 3. Neighbor-joining analysis of *Gossypium gypsy*-like *Gorge1*, *2*, and *3* reverse transcriptase sequences. Unrooted Neighbor-joining analysis of 373 *Gossypium*, 24 *Arabidopsis*, and 36 *Brassica gypsy* reverse transcriptase sequences provides support for the three distinct classes of *gypsy*-like sequences in *Gossypium*. *Gorge1* is similar to *Arabidopsis gypsy* sequence *athila*, *Gorge2* is similar to maize *cinful1*, and *Gorge3* is similar to *del1–46* from *Lilium henryi* and *dea1* from *Ananas comosus*. Bootstrap values for the deeper nodes are shown.

average size of 9.7 kb per *gypsy*, 5.3 kb per *copia*, and 3.5 kb per LINE sequence, we estimate that Class I elements occupy a minimum of 45%–60% of the genome for each of these species, suggesting they have amplified in each lineage approximately in proportion to genome size. However, differential proliferation among species for each group of retrotransposons is evident from the copy number estimates. Copy number estimates for *copia*-like retrotransposons increase proportionally with genome size, with the exception of those from the D genome, which are much higher than expected (Table 2). *Copia*-like sequences occupy 10%–20% of the *G. kirkii*, *G. herbaceum* (A), and *G. exiguum* (K) genome, but have reached considerably higher density in the species with the smallest genome size, *G. raimondii* (D) (28%–39%). LINE-like retroposons are present in similar copy number in the D-genome (13,011; 4.3%) and outgroup *G. kirkii* (16,006; 7.9%) species, but have reached notably higher copy numbers in the A (30,000; 5.3%) and K (27,563; 3.3%) species, both of which contain much larger genomes (Table 2). We were unable to identify SINE-like retroposons (the non-autonomous counterpart of LINEs) in the WGS libraries.

The most striking example of differential lineage-specific amplification of specific groups of repetitive sequences is found among the *gypsy*-like sequences. BLAST analysis led to the discrimination of three different types of *gypsy*-like sequences present in the WGS libraries, and copy-number estimates for each of these are shown separately in Table 2. Phylogenetic analysis of 373 *gypsy*-like reverse transcriptase sequences assembled from all four of the WGS libraries confirmed the existence of these three distinct classes, here designated *Gorge1*, *Gorge2*, and *Gorge3*, for *Gossypium* retrotransposon *gypsy*-like elements (Fig. 2). The *Gorge1* group is similar to the *Arabidopsis gypsy* sequence *athila*, *Gorge2* is similar to maize *cinful1*, and *Gorge3* is similar to *del1–46* from *Lilium henryi* and *dea1* from *Ananas comosus*. Copy number calculations for the three types of sequences revealed relatively stable copy numbers for *Gorge1* and *Gorge2* across all four species, although the copy number estimate for *Gorge1* in the D genome (1971 ± 1762) is somewhat lower than that of the other three species, and *Gorge2* copy number is slightly elevated in the K genome (8220 ± 3983) (Table 2). In contrast to this relative stability for *Gorge1* and *Gorge2*, there is a profound increase in copy number of *Gorge3 gypsy* elements in the larger-genome species. Whereas copy numbers for *Gorge3* are similar in *G. kirkii* and *G. raimondii* (5502 ± 3305 and 8674 ± 3683, respectively), a striking increase in copy number has taken place in both the A (48,181 ± 9257) and K (88,492 ± 12,904) genome lineages. There is a sixfold increase of *Gorge3* copy number from the D to the A genome, and copy number in the A genome is nearly doubled in the K genome. The impact of this proliferation on genome size is apparent from density calculations: *Gorge3* occupies ~9% of both the *G. kirkii* and D genomes, but 27.3% and 33.7% of the A and K genomes, respectively.

Unidentified repetitive fraction

Clones with no similarity to any sequence deposited in GenBank were placed in a separate database as the "unidentified fraction"

of each of the WGS libraries. These sequences were queried against each other using BLASTN to identify repetitive sequences that were missed during the initial BLAST search. Any sequence with >80% sequence to at least three other clones from the same library was considered repetitive. A total of 43, 129, 364, and 603 clones from the *G. kirkii*, D, A, and K libraries, respectively, were considered repetitive under this criterion. The percentage of each library composed of these unidentified repetitive sequences is as follows: *G. kirkii* (OG)—3%; *G. raimondii* (D)—5%; *G. herbaceum* (A)—7.5%; and *G. exiguum* (K)—9%.

Discussion

Variation in nuclear DNA content observed within and between organisms has been a topic of interest dating back to the early 1900s, but was specifically defined and named the "C-value paradox" by Thomas in 1971 (Thomas 1971). Investigations over the past half century have revealed multiple sources of genome size variation, most commonly the differential accumulation or deletion of transposable elements. Repetitive DNA constitutes 80% of angiosperm genomes with haploid DNA content >5.0 pg (Flavell et al. 1974). Approximately 60% or more of the maize (SanMiguel et al. 1996, 1998; Meyers et al. 2001), wheat (Wicker et al. 2001), and barley (Vicient et al. 1999; Shirasu et al. 2000) genomes are composed of transposable elements. Nearly 25% of the maize genome consists of five classes of LTR retrotransposons alone (SanMiguel et al. 1996), and LTR retrotransposon accumulation is responsible for nearly doubling the maize genome in as little as 3 Myr (SanMiguel and Bennetzen 1998). Roughly 80% of the wheat genome is repetitive DNA, mainly LTR retrotransposons (Kumar and Bennetzen 1999). These observations have led to an interest in the effects of repetitive DNA on genome size variation and its significance to plant fitness. Relatively little is known, however, about the evolutionary dynamics of transposable element accumulation among closely related species and how this varies among TE classes.

We constructed WGS libraries for three *Gossypium* and one outgroup species that range approximately fourfold in genome size in order to describe their overall genomic composition and to determine the sequences that contribute to genome size variation. Congruent with results from taxa studied to date, we found that the majority of the *Gossypium* genome consists of dispersed repetitive sequences. Density estimates based on previously reported repetitive sequence lengths suggest that the *Gossypium* genome is composed of ~45%–60% repetitive sequences when considering only those sequences with positive BLAST matches to previously identified repetitive elements in *Gossypium* or in other species. This number is in agreement with estimates from other species with large genomes, such as maize, barley, and wheat (SanMiguel et al. 1996; SanMiguel and Bennetzen 1998; Kumar and Bennetzen 1999; Vicient et al. 1999; Shirasu et al. 2000; Meyers et al. 2001; Wicker et al. 2001), but differs from estimates for smaller-genome species such as *Arabidopsis* (~14%) and rice (~26%) (The *Arabidopsis* Genome Initiative 2000; Jiang and Wessler 2001; Jiang et al. 2004). Additionally, given the number of repetitive sequences of unknown identity recovered in the self-BLAST searches, 45%–60% clearly is an underestimate of the actual repetitive fraction.

Also in agreement with results from other well-studied taxa, the majority of the identified repetitive fraction consists of Class I retrotransposon sequences. As expected based on reported estimates from many grasses and a few well-studied eudicots, Class II sequences were less abundant and constituted a minor fraction of the *Gossypium* genomes (2%). This estimate is comparable to that from *Arabidopsis* (2%–3%) (The *Arabidopsis* Genome Initiative 2000), whose genome is almost five times smaller that that of *G. kirkii* (588 Mb), as well as maize (2%) (Meyers et al. 2001), whose genome is only slightly larger than that of *G. exiguum* (2460 Mb). However, this result is in contrast with that from *Brassica* and rice, whose genomes harbor ~6% and 12% Class II DNA transposons, respectively (Jiang and Wessler 2001; Jiang et al. 2004).

Lineage-specific transposition

A key conclusion of the present study is that genome size variation in a single genus of plants reflects not only the differential amplification of diverse types of repetitive sequences, but that specific families within a repetitive sequence type proliferate differentially as well. From a purely quantitative standpoint, much of the genome size variation observed in *Gossypium* is a consequence of the propagation of one particular family within the larger class of *gypsy*-like retrotransposons, i.e., *Gorge3*. Recently, a *gypsy*-like retrotransposon ("G45" and "G84") that is transcriptionally active was reported in the tetraploid *G. barbadense* (Zaki and Ghany 2004). Comparisons of this active *gypsy* with sequences recovered in the present study revealed a maximum of 96.1% and 80.9% amino acid sequence identity between the best BLAST hit to an A-genome *Gorge3* and G45 and between A-genome *Gorge3* and G84, respectively. Additionally, 165 out of 150,322 *Gossypium* ESTs show >60% sequence similarity over >75 bp to *Gorge3* (e-value cut-off of e-20, data not shown). Based on this high level of sequence identity to G45 and G84, presence in the *Gossypium* EST libraries, and overabundance of *Gorge3* in the WGS libraries, we believe *Gorge3* is a recently active, major constituent of the cotton genome that, like LTR retrotransposons in maize, has triggered a threefold increase in genome size over the 5–10 Myr since the diversification of the major *Gossypium* clades following the origin of the genus (Cronn et al. 2002).

It is interesting to note that other repetitive sequences that are less common than *Gorge3* have also been subject to lineage-specific amplification during diversification of the genus. For example, little amplification of LINE retrotransposons has occurred in the D genome lineage, but these sequences have proliferated in the A and K genome species. Similarly, accumulation of *copia*-like retrotransposons has occurred in the D genome lineage, yet these repetitive elements have been suppressed in the remainder of the genus, with the proportion of the genome occupied by *copias* in the remaining three species being between 10% and 20%. Indeed, *G. raimondii* is the only studied *Gossypium* species in which there are more *copia*-like than *gypsy*-like sequences (Fig. 2).

The most parsimonious interpretation of the *copia* data would invoke differential amplification in the D genome lineage. However, we cannot discount the possibility of unequal rates of DNA loss. Some species appear to be more efficient at removal of non-essential DNA, such that genome size may reflect, at least in part, differential rates of DNA loss (Petrov and Hartl 1997; Kirik et al. 2000; Petrov et al. 2000; Orel and Puchta 2003). With respect to the present study, LINE-like sequences recovered in the WGS libraries are often highly degraded and hence difficult to identify. Although the most parsimonious interpretation of our copy-number estimates is a single amplification event in the common ancestor of the A and K genome lineages, a formal alternative is that LINE-like sequences existed at an ancestrally high copy

number and have subsequently been differentially eliminated from the species with smaller genomes (D genome and the outgroup *G. kirkii*).

Genome size evolution in *Gossypium*

At present relatively little is known about the genomic locations at which genome size evolution takes place in *Gossypium*. The data presented here show that specific families and classes of dispersed repetitive elements have differentially proliferated in different *Gossypium* lineages. Given the propensity of many high-copy-number LTR retrotransposons to accumulate in heterochromatic regions of the genome (Kumar and Bennetzen 1999), we suspect that much of the evolutionarily rapid genome size change that has arisen during the global radiation of *Gossypium* has occurred in these gene-poor regions. Consistent with this notion, Grover et al. (2004) investigated genome size evolution in 104 kb of contiguous sequence surrounding the *CesA1* gene in the *Gossypium* A and D genomes from tetraploid cotton. Within this genic region of the *Gossypium* genome, no evidence of genome size variation was apparent, suggesting that genome size evolution in *Gossypium* takes place in heterochromatic regions located between highly conserved, euchromatic gene islands. Evaluation of this hypothesis will require additional comparative sequence and mapping data, the latter including visualization techniques such as florescent in situ hybridization (FISH) of various transposable elements.

In addition to transposable element accumulation, other suggested mechanisms of genome size change include variation in intron length, expansion/contraction of tandem repeats, illegitimate recombination, indel bias, and unequal intrastrand homologous recombination (Petrov and Wendel 2006). Contrary to suggestions that plants with smaller genomes carry smaller introns (Deutsch and Long 1999; Vinogradov 1999), there is no apparent correlation between genome size and intron length in *Gossypium* (Wendel et al. 2002a; Grover et al. 2004). In fact, intron length has been shown to be highly stable across 28 orthologous sets of genes from A and D genome diploid species and the outgroup species, *G. kirkii* (Wendel et al. 2002a). In the present study, we find no major difference between copy numbers for tandem 5SrDNA and pXP1–80 repeats, although there is a small increase in copy number in larger genomes. However, Cronn et al. (1996) reported a 20-fold variation in 5SrDNA copy number among *Gossypium* species, reflecting both array expansion and contraction. Grover et al. (2004) found no evidence of an indel bias, and although there was some evidence of illegitimate recombination marked by flanking repeats of 2–15 bp in length, the resulting deletions encompass approximately the same proportion of sequence in each genome. Similar studies from other genomic locations in *Gossypium* will be necessary to determine if this result is a local or global occurrence.

Conclusions

Comparative studies of genome size variation among phylogenetically characterized and closely related species serve an important role in clarifying the patterns and processes that underlie the striking genome size variation that characterizes eukaryotes in general and plants in particular. With respect to the latter, we note that the genomic architecture of most plant species remains to be elucidated, and hence mechanisms that characterize one group of plants may not be universal to, say, angiosperms in general. Our data, demonstrating that different families of different classes of TEs have differentially accumulated among closely related clades of a single plant genus, underscores what we believe will be a generality, namely, that mechanisms of genome size evolution are highly variable among even closely related lineages. Our appreciation of plant genomic architecture will continue to be enhanced as comparable studies in other plant groups accumulate. These investigations will generate a deeper understanding of the genomic landscape of different plant lineages, the scale, scope, and pace of evolutionary change responsible for the observed patterns, and insights into the mechanisms the underlie the differential accumulation of different sequence types among genomes.

Methods

Construction and sequencing of WGS libraries

WGS libraries were constructed according to Meyers et al. (2001) with minor modifications and sequenced at the Arizona Genomics Institute, University of Arizona. Briefly, total genomic DNA extracted from young leaves of a single individual was randomly sheared using a Hydroshear (Thorstenson et al. 1998) (GeneMachine), an automated hydrodynamic point-sink–based DNA shearing device (Oefner et al. 1996), at speed code 13 for 25 cycles at room temperature to obtain fragments from *G. herbaceum* (JMS), *G. raimondii* (JFW stock), *G. exiguum* (Gos 5184), and *G. kirkii* (JFW stock) (Fig. 1). Sheared fragments between 2500 and 6000 bp were excised and converted to blunt-ended DNA fragments using the "End-it" DNA end repair kit (Epicentre) containing T4 DNA polymerase (for 5′→3′ polymerase and 3′→5′ exonuclease activities) and T4 Polynucleotide Kinase (for phosphorylation of 5′-ends of blunt DNA), followed by ligation into pBluescriptII KS+ (Strategene) and electroporation into *Escherichia coli* strain DH10B T1 phase-resistant electrocompetent cells (Invitrogen). WGS library clones were sequenced from one direction using the T7 primer (5′-TAATACGACTCACTATAGGG-3′) and BigDye Terminator v3.1 (Applied Biosystems, ABI) according to manufacturer's instruction. Cycle sequencing was performed using PTC-200 thermal cyclers (MJ Research) in a 384-well format with the following regime: 35 cycles of 30 sec at 96°C, 20 sec at 50°C, and 4 min at 60°C. After the cycle-sequencing step, the DNA was purified by ethanol precipitation.

Samples were eluted into 20 µL of water and separated using ABI 3730×l DNA sequencers (ABI). Sequence data were collected and extracted using sequence analysis software (ABI). The sequencing data were base-called using the program Phred (Ewing et al. 1998). Vector and low-quality sequences were removed by the program Lucy (Chou and Holmes 2001) and then submitted to the GSS division of GenBank under accessions DX390732–DX406528.

Analytical framework

The number of sequences needed to generate 95% confidence that at least one member of a given class of sequences will be sampled was determined for each species using the following equation:

$$N_{.95} = \ln (0.05) / \ln \{1 - [n (l - 2m + e)/(G - e)]\} \qquad (1)$$

where $N_{.95}$ is the sampling effort required to be 95% confident that at least one target sequence will be sampled, n is the number of targets present in the genome, l is the length of the target sequence, m is the estimated minimum length required to identify the sequence in a BLAST search, e is the number of base pairs sequenced from each insert, and G is genome size. By using this

equation we were able to estimate the sampling intensity needed to detect at least one repetitive sequence of an estimated length and copy number in each of the four genomes. Published data for diverse types of repetitive elements, such as *Ty3-gypsy*, *Ty1-copia*, LINE retroposons, SINEs, and MITEs of various estimated lengths (l) and copy numbers (n) were used to calculate $N_{.95}$ in order to determine how many clones should be sequenced from each library. The value for m was conservatively estimated at $m = 200$ bp, which in a BLASTX analysis would equal sequence similarity over ≥66 amino acids. We estimated e to be 700 bp, based on the average high-quality sequencing read length reported from the Arizona Genomics Institute. Based on these estimates, libraries were constructed that contain 1.5% (based on ~5 kb plasmid insert length) of the genome from each species. One-pass sequencing from one end of the insert (e ~700 bp) was performed, which, when totaled across the number of clones sequenced, yielded sequence data for ~0.2% of each haploid genome (Table 1).

Data analysis and copy number estimation

Because of rapid sequence divergence of repetitive DNA and the limited database of repetitive sequences available in GenBank for plants closely related to *Gossypium*, sequences from the WGS libraries were subjected to BLASTX (amino acid) in addition to BLASTN (nucleotide) analyses at the NCBI Web site. Hits of e-5 or better were retained for further analysis. In addition, libraries were queried against themselves in an attempt to identify families of repetitive elements not recognized in the initial search. In this self-BLAST analysis, sequences with >80% identity over 100 bp were considered related. Clones were assigned to a general category according to their best BLAST hit. These general categories were (1) nuclear, (2) chloroplast, (3) mitochondrial, (4) repetitive, and (5) unknown.

Plant transposable elements are broadly divided into three main lineages: the "Transposons" consisting of the Class II DNA elements, the "Retrotransposons" containing the LTR Class I elements, and the "Retroposons" consisting of the non-LTR Class I elements (Eickbush and Malik 2002). Class I elements transpose via a duplicative mechanism, in which an RNA intermediate formed from the parental copy is reverse transcribed, and the newly translated copies are inserted into new positions in the genome. Class I LTR retrotransposons are subdivided into two classes, *gypsy* and *copia*-like, based on the position of the integrase coding domain. The non-LTR retroposons consist of autonomous Long Interspersed Nuclear Elements (LINEs) and the non-autonomous Small Interspersed Nuclear Elements (SINEs). Class II DNA elements transpose via a cut-and-paste mechanism in which the element is excised and inserted into a new area of the genome. DNA elements characteristically contain terminal inverted repeats (TIR) ranging from 11 to a few hundred base pairs in length, and families of elements are defined by these TIR sequences (Bennetzen 2000b). Class II transposons can be divided into three main superfamilies: *hAT* (*hobo* from *Drosophila*, *Activator* of maize, and *Tam* from Snapdragon), *Mutator*, and *En/Spm*, both first described in maize (Kidwell 2002). Therefore, dispersed repetitive sequences recovered from the WGS libraries were placed into the specific categories 1) *gypsy*-like, 2) *copia*-like, 3) LINE-like, 4) *hAT*-like, 5) *En/Spm*-like, and 6) *Mutator*-like. DNA sequence alignments were performed with published sequences of the same type to confirm sequence identity.

Tandem repeats were identified using the program Tandem Repeat Finder (Benson 1999). Searches were performed using the default settings. Any tandem repeat present in more than three clones with a score >500 was retained for further analysis. These sequences were queried against GenBank using BLASTN to search for sequence similarity to known sequences deposited in GenBank. Sequences were queried against one another to identify sequences that were shared among the libraries.

Copy numbers (n) for various repetitive elements recovered from the WGS libraries were estimated according to the following equation:

$$n = (X_{obs} / N)(G - e)(1 / (l - 2m + e)) \quad (2)$$

where X_{obs} is the observed number of copies, N is the total number of sequence reads, and the other variables are as before: n = number of targets in the genome; l = length of target sequence; m = estimated minimum length required to identify sequence in a BLAST search; e = number of bp sequenced from each insert; and G = genome size. Published sequences for various repetitive elements were used to estimate l. Similar to average *copia* sequences in rice (5–6 kb) (McCarthy et al. 2002), an l of 5.3 kb was used for *copia*-like sequences based on published data from *Gossypium* (Grover et al. 2004). Also in agreement with rice data for *gypsy*-like sequences (11–13 kb) (McCarthy et al. 2002), l for gypsy-like sequences in *Gossypium* was set at 9.7 kb (C.E. Grover, unpubl.). Because no data exist for other dispersed repetitive sequences in *Gossypium*, the estimated lengths for the remaining repetitive sequences were established according to their closest BLAST hit from GenBank and are assigned as follows: LINE retroposon 3.5 kb (GenBank accession no. NP_92230 from *O. sativa*); *En/Spm* with high identity to *Tam1* (Nacken et al. 1991) (GenBank accession no. X57297) 15.2 kb; and *hAT* with high identity to *Tam3* (Hehl et al. 1991) (GenBank accession no. X55078) 3.6 kb. Published lengths for *Mutator*-like sequences are highly variable (in *Arabidopsis* these range from 444 to 19,397 bp) (Yu et al. 2002); therefore, we did not attempt to estimate *Mutator*-like copy numbers. A recent manuscript by Rabinowicz et al. (2005) used WGS libraries to estimate gene number in various plant species. When using their data for *Arabidopsis* and rice in our equation, we recover comparable results to those published for these two sequenced genomes, suggesting that our equation results in reasonably accurate estimates of copy numbers.

Phylogeny reconstruction

Sequences were queried against coding domains of various repetitive sequences from *Arabidopsis thaliana* and *Brassica oleracea* obtained from S. Wessler and F. Zhang (Univ. Georgia). Amino acid sequences with an e-value of e-5 or better were imported into BioEdit (Hall 1999) and aligned using ClustalW (Johnson et al. 1994). Neighbor-Joining analysis was performed in Paup* (Swofford 2001) using the default settings.

Acknowledgments

We thank Curt Brubaker for providing *Gossypium exiguum* seeds and DNA; S. Wessler and F. Zhang for sharing a repetitive sequence database; Jordan Swanson for assistance with data analysis; Ryan Percifield, Corrinne Grover, and Ryan Rapp for helpful discussion; and the anonymous reviewers for their comments. This work was funded by the National Science Foundation Plant Genome program.

References

Adams, K.L. and Palmer, J.D. 2003. Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Mol. Phylogenet. Evol.* **29:** 380–395.

Ananiev, E.V., Phillips, R.L., and Rines, H.W. 1998. Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. *Proc. Natl. Acad. Sci.* **95:** 13073–13078.
The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.
Bancroft, I. 2001. Duplicate and diverge: The evolution of plant genome microstructure. *Trends Genet.* **17:** 89–93.
Beasley, J.O. 1941. Hybridization, cytology, and polyploidy of *Gossypium*. *Chron. Bot.* **6:** 394–395.
Bennett, M.D. and Leitch, I.J. 1995. Nuclear DNA amounts in angiosperms. *Ann. Bot. (Lond.)* **76:** 113–176.
———. 1997. Nuclear DNA amount in angiosperms. *Philos. Trans. R. Soc. Lond., B Biol. Sci.* **334:** 309–345.
———. 2005. Plant genome size research: A field in focus. *Ann. Bot. (Lond.)* **95:** 1–6.
Bennett, M.D. and Smith, J.B. 1991. Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond., B Biol. Sci.* **334:** 309–345.
Bennetzen, J.L. 2000a. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell* **12:** 1021–1029.
———. 2000b. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42:** 251–269.
———. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* **115:** 29–36.
Bennetzen, J.L. and Kellogg, E.A. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9:** 1509–1514.
Bennetzen, J.L. and Ramakrishna, W. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol. Biol.* **48:** 821–827.
Bennetzen, J.L., Ma, J., and Devos, K.M. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond.)* **95:** 127–132.
Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.
Biderre, C., Metenier, G., and Vivares, C.P. 1998. A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi*. *Mol. Biochem. Parasitol.* **74:** 229–231.
Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12:** 1093–1101.
Cavalier-Smith, T. 1985. *The evolution of genome size*. John Wiley, New York.
Chen, M., SanMiguel, P., and Bennetzen, J.L. 1998. Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. *Genetics* **148:** 435–443.
Chooi, W.Y. 1971. Variation in nuclear DNA content in the genus *Vicia*. *Genetics* **68:** 195–211.
Chou, H.-H. and Holmes, M.H. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics* **17:** 1093–1104.
Cronn, R.C., Zhao, X., Paterson, A.H., and Wendel, J.F. 1996. Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. *J. Mol. Evol.* **42:** 685–705.
Cronn, R.C., Small, R.L., Haselkorn, T., and Wendel, J.F. 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *Am. J. Bot.* **89:** 707–725.
Deutsch, M. and Long, M. 1999. Intron–exon structure of eukaryotic model organisms. *Nucleic Acids Res.* **27:** 3219–3228.
Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12:** 1075–1079.
Eickbush, T.H. and Malik, H.S. 2002. Origin and evolution of retrotransposons. In *Mobile DNA II* (eds. N.L. Craig et al.), pp. 1111–1144. ASM Press, Washington, D.C.
Ellegren, H. 2002. Mismatch repair and mutational bias in microsatellite DNA. *Trends Genet.* **18:** 552.
Endrizzi, J.E., Turcotte, E.L., and Kohel, R.J. 1985. Genetics, cytogenetics, and evolution of *Gossypium*. *Adv. Genet.* **23:** 271–375.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequences traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.
Filkowski, J., Kovalchuk, O., and Kovalchuk, I. 2004. Dissimilar mutation and recombination rates in *Arabidopsis* and tobacco. *Chromosoma* **166:** 265–272.
Flavell, R.B., Bennett, M.D., Smith, J.B., and Smith, D.B. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* **12:** 257–269.
Fryxell, P.A. 1992. A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). *Rheedea* **2:** 108–165.
Grant, D., Cregan, P., and Shoemaker, R.C. 2000. Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **97:** 4168–4173.
Gregory, T.R. 2002. A bird's eye view of the C-value enigma: Genome size, cell size, and metabolic rate in the class Aves. *Evolution Int. J. Org. Evolution* **56:** 121–130.
———. 2004. Macroevolution, hierarchy theory, and the C-value enigma. *Paleobiology* **30:** 179–202.
Grover, C.E., Kim, H., Wing, R.A., Paterson, A.H., and Wendel, J.F. 2004. Incongruent patterns of local and global genome size evolution in cotton. *Genome Res.* **14:** 1474–1482.
Hall, T.A. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acid Symp. Ser.* **41:** 95–98.
Hall, S.E., Kettler, G., and Preuss, D. 2003. Centromere satellites from *Arabidopsis* populations: Maintenance of conserved and variable domains. *Genome Res.* **13:** 195–205.
Hehl, R., Nacken, W.K., Krause, A., Saedler, H., and Sommer, H. 1991. Structural analysis of Tam3, a transposable element from *Antirrhinum majus*, reveals homologies to the Ac element from maize. *Plant Mol. Biol.* **16:** 369–371.
Hendrix, B. and Stewart, J.M. 2005. Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot. (Lond.)* **95:** 789–797.
Ito, H., Nasuda, S., and Endo, T.R. 2004. A direct repeat sequence associated with the centromeric retrotransposons in wheat. *Genome* **47:** 747–756.
Jiang, N. and Wessler, S.R. 2001. Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell* **13:** 2533–2564.
Jiang, N., Feschotte, C., Zhang, X., and Wessler, S.R. 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.* **7:** 115–119.
Johnson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Jones, R.N. and Brown, L.M. 1976. Chromosome evolution and DNA variation in *Crepsis*. *Heredity* **36:** 91–104.
Kidwell, M.G. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115:** 49–63.
Kirik, A., Salomon, S., and Puchta, H. 2000. Species-specific double-strand break repair and genome evolution in plants. *EMBO J.* **2000:** 5562–5566.
Ku, H.-M., Vision, T., Liu, J., and Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci.* **97:** 9121–9126.
Kumar, A. and Bennetzen, J.L. 1999. Plant retrotransposons. *Annu. Rev. Genet.* **33:** 479–532.
Leitch, I.J., Chase, M.W., and Bennett, M.D. 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Annals Bot. (Suppl. A)* **82:** 85–94.
Li, W.-H. and Graur, D. 1991. *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, MA.
Ma, J., Devos, K.M., and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14:** 860–869.
McCarthy, E.M., Liu, J., Lizhi, G., and McDonald, J.F. 2002. Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol.* **3:** 1–11.
Meyers, B.C., Tingey, S.V., and Morgante, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11:** 1660–1676.
Morgante, M., Hanafey, M., and Powell, W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30:** 194–200.
Nacken, W.K., Piotrowiak, R., Saedler, H., and Sommer, H. 1991. The transposable element Tam1 from *Antirrhinum majus* shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol. Genet. Genomics* **228:** 201–208.
Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C.R., Cheng, Z., and Jiang, J. 2004. Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol. Biol. Evol.* **4:** 845–855.
Oefner, P.J., Hunicke-Smith, S.P., Chiang, L., Dietrich, F., Mulligan, J., and Davis, R.W. 1996. Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res.* **24:** 3879–3886.
Orel, N. and Puchta, H. 2003. Differences in the processing of DNA ends

in *Arabidopsis thaliana* and tobacco: Possible implications for genome evolution. *Plant Mol. Biol.* **51:** 523–531.
Petrov, D.A. 2002a. DNA loss and evolution of genome size in *Drosophila*. *Genetica* **115:** 81–91.
———. 2002b. Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61:** 531–544.
Petrov, D.A. and Hartl, D.L. 1997. Trash DNA is what gets thrown away: High rate of DNA loss in *Drosophila*. *Gene* **205:** 279–289.
Petrov, D.A. and Wendel, J.F. 2006. Genome evolution in eukaryotes: The genome size perspective. In *Evolutionary genetics: Concepts and case studies* (eds. C.W. Fox and J.B. Wolf), pp. 144–156. Oxford University Press, Oxford, UK.
Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. 2000. Evidence for DNA loss as a deteminant of genome size. *Science* **287:** 1060–1062.
Price, H.J. 1988. Nuclear DNA content variation within angiosperm species. *Evol. Trends Plants* **2:** 53–60.
Rabinowicz, P.D., Citek, R., Budiman, M.A., Nunberg, A., Bedell, J.A., Lakey, N., O'Shaughnessy, A.L., Nascimento, L.U., McCombie, W.R., and Martienssen, R.A. 2005. Differential methylation of genes and repeats in land plants. *Genome Res.* **15:** 1431–1440.
SanMiguel, P. and Bennetzen, J.L. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot. (Lond.)* **82:** 37–44.
SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.
SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.
Seelanan, T., Schnabel, A., and Wendel, J.F. 1997. Congruence and consensus in the cotton tribe. *Syst. Bot.* **22:** 259–290.
Shahmuradov, I.A., Akbarova, Y.Y., Solovyev, V.V., and Aliyev, J.A. 2003. Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *Plant Mol. Biol.* **52:** 923–934.
Shepherd, N.S., Schwarz-Sommer, Z., Blumberg vel Spalve, J., Gupta, M., Wienand, U., and Saidler, H. 1984. Similarity of the Cin1 repetitive family of *Zea mays* to eukaryotic transposable elements. *Nature* **307:** 185–187.
Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10:** 908–915.
Small, R.L., Ryburn, J.A., Cronn, R.C., Seelanan, T., and Wendel, J.F. 1998. The tortoise and the hare: Choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. *Am. J. Bot.* **85:** 1301–1315.
Small, R.L., Ryburn, J.A., and Wendel, J.F. 1999. Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). *Mol. Biol. Evol.* **16:** 491–501.
Sparrow, A.H., Price, H.J., and Underbrink, A.G. 1972. A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: Some evolutionary considerations. *Brookhaven Symp. Biol.* **23:** 451–494.
Swofford, D.L. 2001. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA.
Thomas, C.A. 1971. The genetic organisation of chromosomes. *Annu. Rev. Genet.* **5:** 237–256.
Thorstenson, T.R., Hunicke-Smith, S.P., Oefner, P.J., and Davis, R.W. 1998. An automated hydrodynamic process for controlled, unbiased DNA shearing. *Genome Res.* **8:** 848–855.
Tikhonov, A., SanMiguel, P., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc. Natl. Acad. Sci.* **96:** 7409–7414.
Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., and Schulman, A.H. 1999. Retrotransposon *BARE*-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell* **11:** 1769–1784.
Vinogradov, A.E. 1999. Intron–genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49:** 376–384.
Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290:** 2114–2117.
Wendel, J.F. 2000. Genome evolution in polyploids. *Plant Mol. Biol.* **42:** 225–249.
Wendel, J.F. and Albert, V.A. 1992. Phylogenetics of the cotton genus (*Gossypium*): Character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Syst. Bot.* **17:** 115–143.
Wendel, J.F. and Cronn, R.C. 2003. Polyploidy and the evolutionary history of cotton. *Adv. Agron.* **78:** 139–186.
Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L., and Senchina, D.S. 2002a. Intron size and genome size in plants. *Mol. Biol. Evol.* **19:** 2346–2352.
Wendel, J.F., Cronn, R.C., Johnston, J.S., and Price, H.J. 2002b. Feast and famine in plant genomes. *Genetica* **115:** 37–47.
Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.* **26:** 307–316.
Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.-D., Dubcovsky, J., and Keller, B. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell* **15:** 1186–1197.
Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79–92.
Zaki, E.A. and Ghany, A.A.A. 2004. *Ty3/gypsy* retro-transposons in Egyptian cotton (*G. barbadense*). *J. Cotton. Sci.* **8:** 179–185.
Zhang, J. 2003. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **18:** 292–298.
Zhao, X.P., Si, Y., Hanson, R.E., Crane, C.F., Price, H.J., Stelly, D.M., Wendel, J.F., and Paterson, A.H. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res.* **8:** 479–492.

Received March 3, 2006; accepted in revised form May 22, 2006.

Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*

Jennifer S. Hawkins, HyeRan Kim, John D. Nason, et al.

Genome Res. 2006 16: 1252-1261
Access the most recent version at doi:10.1101/gr.5282906